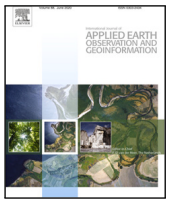
International Journal of Applied Earth Observation and Geoinformation 143 (2025) 104817

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



AeroReformer: Aerial Referring Transformer for UAV-based Referring Image Segmentation

Rui Li, Xiaowei Zhao *

Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry, CV4 7AL, UK

ARTICLE INFO

Dataset link: https://uavid.nl/, https://huggingface.co/datasets/RussRobin/VDD, https://github.com/lironui/AeroReformer, https://huggingface.co/datasets/lironui/UAVid-RIS, https://huggingface.co/datasets/lironui/VDD-RIS

Keywords:
UAV-RIS
MLLM
Referring image segmentation
Deep learning
UAV

ABSTRACT

As a novel and challenging task, referring segmentation combines computer vision and natural language processing to localise and segment objects based on textual descriptions. While Referring Image Segmentation (RIS) has been extensively studied in natural images, little attention has been given to aerial imagery, particularly from Unmanned Aerial Vehicles (UAVs). The unique challenges of UAV imagery, including complex spatial scales, occlusions, and varying object orientations, render existing RIS approaches ineffective. A key limitation has been the lack of UAV-specific datasets, as manually annotating pixel-level masks and generating textual descriptions is labour-intensive and time-consuming. To address this gap, we design an automatic labelling pipeline that leverages pre-existing UAV segmentation datasets and the Multimodal Large Language Models (MLLM) for generating textual descriptions. Furthermore, we propose Aerial Referring Transformer (AeroReformer), a novel framework for UAV Referring Image Segmentation (UAV-RIS), featuring a Vision-Language Cross-Attention Module (VLCAM) for effective cross-modal understanding and a Rotation-Aware Multi-Scale Fusion (RAMSF) decoder to enhance segmentation accuracy in aerial scenes. Extensive experiments on two newly developed datasets demonstrate the superiority of AeroReformer over existing methods, establishing a new benchmark for UAV-RIS. The datasets and code are publicly available at https://github.com/lironui/AeroReformer.

1. Introduction

Referring Image Segmentation (RIS) aims to segment target objects in an image based on natural language expressions that describe their attributes or context (Li et al., 2018; Ding et al., 2022). Unlike traditional image segmentation methods that rely on predefined semantic labels and operate within a constrained set of categories (Simonyan and Zisserman, 2014; Ronneberger et al., 2015; Wang et al., 2022), referring image segmentation enables open-domain segmentation by utilising free-form textual descriptions as guidance (Hu et al., 2016; Liu et al., 2017; Lai et al., 2024). This capability significantly expands its applicability, allowing for more flexible and context-aware interpretation of imagery. In terms of the aerial scenario, UAV Referring Image Segmentation (UAV-RIS) has broad applications in domains such as text-guided environmental monitoring (Sharma and Arya, 2022), land cover classification (Mienna et al., 2022), precision agriculture (Tahir et al., 2023), urban planning (Shao et al., 2021) and risk assessment (Trepekli et al., 2022), where identifying and segmenting specific objects or regions based on natural language descriptions is crucial. By leveraging the multimodal integration of vision and language, UAV-RIS can enhance the precision and adaptability of spatial analysis, making it easier to extract detailed, context-specific information from complex aerial imagery (see Fig. 1).

Recently, benefiting from the open-source datasets including RefSegRS (Yuan et al., 2024), RRSIS-D (Liu et al., 2024) and RISBench (Dong et al., 2024), the Referring Remote Sensing Image Segmentation (RRSIS) has attracted more and more attention (Lei et al., 2024; Shi and Zhang, 2025; Zhang et al., 2025; Chen et al., 2025; Li et al., 2025). Despite these promising advances, UAV-RIS poses additional complexities due to the lower altitudes and agile motion of UAV platforms, leading to more pronounced occlusions, rapidly shifting viewpoints, and varying scene contexts (Lyu et al., 2020; Li and Zhao, 2024; Zhang et al., 2023). Moreover, building a large-scale, high-quality UAV dataset for referring segmentation remains labour-intensive, underscoring the need for automated or semi-automated approaches to annotation. Therefore, it remains under-investigated in aerial imagery, particularly for data captured by UAVs.

This paper addresses the challenges of UAV-RIS by introducing UAV-specific datasets and a novel framework, expanding the scope of referring image segmentation to UAV imagery and establishing

* Corresponding author.
E-mail addresses: rui.li.8@warwick.ac.uk (R. Li), xiaowei.zhao@warwick.ac.uk (X. Zhao).

https://doi.org/10.1016/j.jag.2025.104817
Received 28 February 2025; Received in revised form 20 August 2025; Accepted 23 August 2025

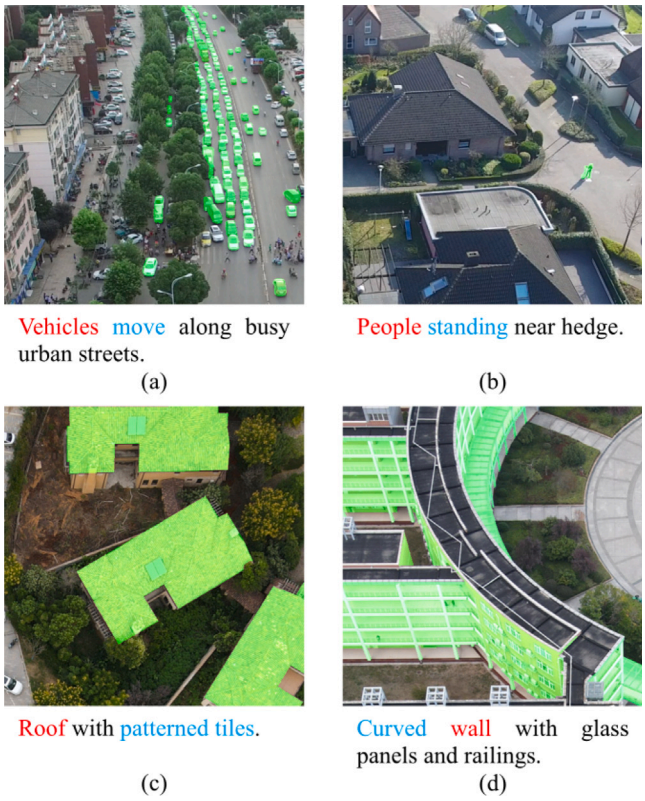


Fig. 1. The shifting viewpoints and varying scene contexts are very common for UAV-captured images.

a foundation for future research. Specifically, we develop an automatic labelling pipeline that leverages open-source and pre-existing UAV segmentation datasets along with a Multimodal Large Language Model (MLLM). In our data generation process, segmentation masks are obtained from existing dataset annotations, and a cropped image paired with a well-designed prompt is fed into the MLLM to generate textual descriptions of the target object. This approach streamlines the annotation and description generation process, reducing the time and effort required for manual labelling.

Meanwhile, UAV imagery presents unique challenges compared to natural images, including significant scale variations, diverse object orientations, and complex background clutter (Lyu et al., 2020; Li and Zhao, 2024; Zhang et al., 2023). To effectively bridge the gap between visual and linguistic modalities, we propose a UAV-specific RIS model, *i.e.* AeroReformer, featuring a Vision-Language Cross-Attention Module (VLCAM) for robust cross-modal understanding and a Rotation-Aware Multi-Scale Fusion (RAMSF) decoder to address spatial variations in UAV imagery. VLCAM dynamically aligns visual features with linguistic queries, ensuring that textual descriptions are accurately mapped to corresponding image regions, even under complex conditions such as occlusions and scale variations. Meanwhile, RAMSF enhances the segmentation process by incorporating rotation-aware convolutions and multi-scale feature aggregation, preserving orientation consistency while maintaining high-resolution spatial details. The integration of these two modules enables our model to outperform existing methods, achieving state-of-the-art results on UAVid-RIS and VDD-RIS, two datasets generated by our proposed pipeline. The main contribution of this paper can be summarised as:

(1) An automatic dataset generation framework is designed, enabling the transformation of labelled segmentation datasets into their LLM-aided counterparts.
(2) Two UAV-RIS datasets, UAVid-RIS and VDD-RIS, are constructed from open-source datasets, providing a benchmark for UAV-RIS research and evaluation.
(3) A novel UAV-RIS network, AeroReformer, is designed, incorporating a Vision-Language Cross-Attention Module (VLCAM) and a Rotation-Aware Multi-Scale Fusion (RAMSF) decoder, achieving state-of-the-art performance on UAVid-RIS and VDD-RIS.

The remainder of this paper is organised as follows: Section 2 reviews related work. Section 3 presents the UAV referring segmentation dataset generation pipeline. Section 4 introduces the proposed AeroReformer model, explaining its architecture. Section 5 describes the experimental setup, dataset details, and evaluation metrics and presents a performance analysis. Finally, Section 6 concludes the research and discusses potential directions for future research.

2. Related work

2.1. Referring image segmentation for natural images

RIS is a fundamental task in vision-language understanding, where the goal is to segment objects in an image based on a given natural language expression (Li et al., 2018; Yu et al., 2016, 2018). This task demands a fine-grained alignment between textual descriptions and visual features to correctly localise and delineate the referenced objects. Compared to conventional segmentation tasks that rely on predefined categories, RIS enables a more flexible and user-specific segmentation process.

In the early stages, initial RIS models relied primarily on Convolutional Neural Networks (CNNs) to extract visual features and Recurrent Neural Networks (RNNs) to process textual descriptions (Li et al., 2018; Hu et al., 2016; Nagaraja et al., 2016). These models performed feature fusion by concatenating visual and linguistic representations before feeding them into a segmentation head. Specifically, Hu et al. (2016) first introduced RIS to address the limitations of traditional semantic segmentation when handling complex textual descriptions. Later, Li et al. (2018) and Nagaraja et al. (2016) explored bidirectional interactions between visual and textual features, improving the multimodal understanding of objects through structured representations. Further advancements introduced dynamic multimodal networks, such as the work by Margffoy-Tuay et al. (2018), which incorporated recursive reasoning mechanisms to enhance the integration of linguistic and visual information.

As RIS models evolved, researchers recognised the importance of cross-modal feature alignment, leading to the introduction of attention-based strategies (Shi et al., 2018; Ye et al., 2019; Hu et al., 2020). For example, Shi et al. (2018) introduced a keyword-aware segmentation model, refining object-region relationships based on key linguistic cues. These approaches significantly improved object localisation and contextual interpretation in RIS tasks. Ye et al. (2019) proposed a cross-modal self-attention module to capture long-range dependencies between textual and visual elements, improving multimodal fusion. Similarly, Hu et al. (2020) developed a bidirectional cross-modal attention mechanism, enabling deeper interaction between the modalities.

The recent emergence of Transformer-based architectures has significantly advanced RIS, offering global modelling capabilities and superior multimodal integration. Unlike CNN-based methods, which rely on local receptive fields, Transformers enable long-range dependencies and self-attention mechanisms, making them particularly effective for RIS (Ding et al., 2022; Yang et al., 2022; Liu et al., 2023). Several notable works have leveraged this architecture. VLT designed a query-based Transformer framework, enriching textual comprehension by dynamically generating language query embeddings (Ding et al., 2022). LAVT proposed language-aware attention mechanisms to enhance early fusion between the two modalities, enabling more precise segmentation (Yang et al., 2022). GRES further refined multimodal alignment by explicitly modelling dependencies between different textual tokens and visual regions, leading to more robust segmentation performance (Liu et al., 2023).

2.2. Referring remote sensing image segmentation

Referring Remote Sensing Image Segmentation (RRSIS) is a specialised task that aims to extract pixel-wise segmentation masks from remote sensing imagery based on natural language expressions (Yuan et al., 2024; Liu et al., 2024). While it has significant applications in environmental monitoring, land cover classification, disaster response, and urban planning (Sun et al., 2022; Li et al., 2024), progress in this field hinges critically on suitable datasets that capture the complexity of remote sensing imagery. One of the earliest datasets was RefSegRS, introduced in Yuan et al. (2024), which enabled initial efforts to adapt RIS methods from natural images to the remote sensing domain. To enhance the diversity and improve the generalisability of trained models, Liu et al. (2024) proposed RRSIS-D, a substantially larger dataset for benchmarking mainstream RIS models in remote sensing image segmentation. More recently, RISBench (Dong et al., 2024) has also been introduced to further advance the development and evaluation of RRSIS methods.

Building on these datasets, recent RRSIS research has explored strategies to address scale variations, complex backgrounds, and orientation diversity. LGCE (Yuan et al., 2024) pioneered a language-guided cross-scale enhancement module to fuse shallow and deep features for improved segmentation accuracy, whereas (Liu et al., 2024) proposed the Rotated Multi-Scale Interaction Network (RMSIN), which integrates intra-scale and cross-scale interactions alongside rotated convolutions to better handle directional variations. Beyond scale-aware models, Pan et al. (2024) analysed the implicit optimisation mechanisms in existing models and proposed an explicit affinity alignment approach, incorporating a new loss function to improve textual-visual feature interaction. More recent studies have introduced refined image-text alignment strategies to improve RRSIS performance. Specifically, FIANet (Lei et al., 2024) introduced a fine-grained alignment module with object-positional enhancement, integrating a text-aware self-attention mechanism to refine segmentation accuracy. Similarly, CroBIM (Dong et al., 2024) leveraged a context-aware prompt modulation module, which optimises post-fusion feature interactions and employs a mutual-interaction decoder to refine segmentation masks. Recently, SBANet (Li et al., 2025) introduced a bidirectional alignment mechanism and a scale-wise attention module to enhance mutual guidance between vision and language features, effectively refining segmentation masks in referring remote sensing image segmentation. BTDNet (Zhang et al., 2025) employs a bidirectional spatial correlation module and a target-background twin-stream decoder to improve multimodal alignment and fine-grained object differentiation, achieving improved segmentation performance.

2.3. Visual grounding for aerial images

Another active vision-and-language research in the remote sensing community is visual grounding for aerial images, focusing on localising target objects within aerial scenes using natural language queries (Sun et al., 2022; Zhao et al., 2021; Zhan et al., 2023). In contrast to RRSIS, which demands detailed pixel-level masks, visual grounding is primarily concerned with identifying object-level regions, typically represented as bounding boxes (Sun et al., 2022). This task leverages the unique characteristics of aerial imagery, where targets often exhibit complex spatial relationships and may not be visually prominent due to scale variations and cluttered backgrounds.

Early frameworks, such as GeoVG (Sun et al., 2022), pioneered this approach by integrating a language encoder that captures geospatial relationships with an image encoder that adaptively attends to aerial scenes. By fusing these modalities, GeoVG established a one-stage process that effectively translates natural language cues into object localisation. Building on this foundation, subsequent models have introduced advanced fusion strategies. For instance, modules like the Transformer-based Multi-Granularity Visual Language Fusion (MGVLF) (Zhan et al., 2023) exploit both multi-scale visual features and multi-granularity textual embeddings, resulting in more discriminative representations that address the challenges posed by large-scale variations and busy backgrounds. Vision-Semantic Multimodal Representation (VSMR) enhanced multi-level feature integration, refining how visual and textual features are jointly processed to improve localisation robustness (Ding et al., 2024). Further improvements have been achieved through progressive attention mechanisms. The Language-guided Progressive Visual Attention (LPVA) framework, for example, dynamically adjusts visual features at various scales and levels, ensuring that the visual backbone concentrates on expression-relevant information (Li et al., 2024). This is complemented by multi-level feature enhancement decoders, which aggregate contextual information to boost feature distinctiveness and suppress irrelevant regions.

3. UAV referring segmentation dataset generation

Although several RRSIS datasets have been introduced, they predominantly focus on vertically captured (nadir-view) satellites and aerial imagery, limiting their applicability to UAV-based scenarios. Unlike satellite imagery, UAVs operate at lower altitudes with dynamic viewing angles, resulting in significant variations in object appearance due to oblique perspectives, occlusions, and scale distortions. These factors make existing RRSIS datasets insufficient for UAV-RIS tasks, where diverse viewpoints and fine-grained scene details are crucial for accurate segmentation. To address this gap, we introduce a UAV-RIS dataset generation pipeline, ensuring more realistic and comprehensive benchmarking for UAV-based vision-language tasks.

3.1. Dataset construction and analysis

To advance UAV-RIS, we present a fully automated pipeline that leverages pre-existing UAV segmentation datasets. Unlike traditional approaches requiring manual annotations, our pipeline efficiently generates language-vision pairs by integrating segmentation masks with large language models. The data generation process, as shown in Fig. 2, is structured as follows:

- **Step 1: Segmentation-based Cropping.**
  Given a pre-labelled UAV segmentation dataset, images and their corresponding masks are first partitioned into 1024×1024 patches. To ensure meaningful segmentation patches, we apply a filtering mechanism based on class presence and distribution. For each patch, the class distribution is computed by analysing the pixel proportions of predefined categories such as buildings, trees, roads, and vehicles. Patches containing only a single dominant class (occupying more than 70% of the patch) or those with minimal class representation (below predefined thresholds) are discarded. This step enhances dataset diversity and ensures the presence of multiple meaningful classes in each selected patch.
- **Step 2: Vision-Language Description Generation.**
  Once cropped patches are obtained, the segmented regions are processed using Qwen2.5-VL-7B (Bai et al., 2025), a vision-language model capable of generating concise and context-aware descriptions for detected objects. The model takes an image patch as input, along with a predefined prompt specifying the target object class. Each prompt ensures the inclusion of the relevant object category while avoiding semantically conflicting terms. The object categories vary depending on the pre-labelled segmentation dataset used. Each dataset's predefined classes are utilised to generate corresponding descriptions, ensuring alignment with its original annotations. Tailored instructions are provided for each category to produce accurate and concise textual descriptions. To maintain consistency, the model operates under a controlled generation setting where responses are constrained to a maximum of 10 words. Additionally, the prompt explicitly instructs the

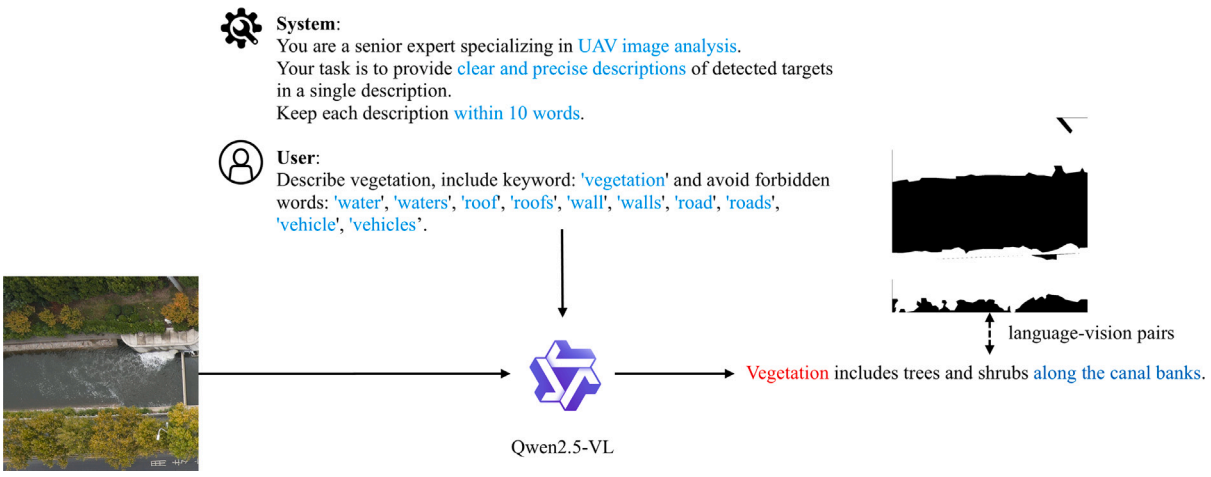
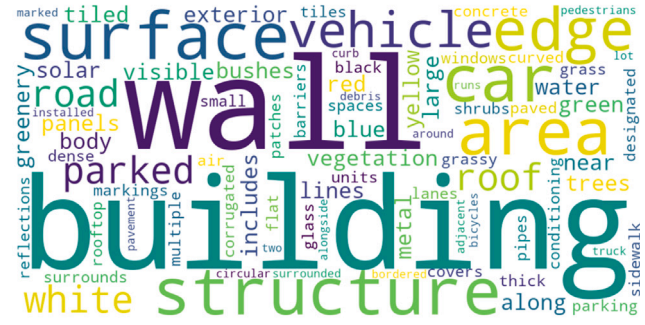


Fig. 2. The automatic UAV referring segmentation dataset generation pipeline.

(a) Word cloud of top 100 words in UAVid-RIS



(b) Word cloud of top 100 words in VDD-RIS

Fig. 3. Word cloud for top 100 words within the expressions of (a) UAVid-RIS and (b) VDD-RIS.

model to avoid irrelevant or misleading terms, ensuring that the generated descriptions remain semantically aligned with the visual content. The processed text-image pairs form the basis of the final dataset, enabling high-quality referring segmentation in UAV imagery.

- **Step 3: Automatic Description Refinement.**
  To maintain dataset consistency and clarity, we implement a text-cleaning process that removes ambiguous or uninformative phrases, such as *"no visible"*. This automatic post-processing step ensures that all descriptions remain meaningful and directly correspond to the visual content of the segmented region. Finally, the annotations are formatted to align with the RefCOCO (Lin et al., 2014) dataset structure, enhancing compatibility with existing referring segmentation models and facilitating seamless integration into vision-language benchmarks.

Our proposed pipeline has been applied to two widely used UAV segmentation datasets, UAVid (Lyu et al., 2020) and VDD (Cai et al., 2023), to generate their corresponding referring segmentation versions, namely UAVid-RIS and VDD-RIS. By leveraging the pre-existing segmentation annotations, our approach automatically extracts meaningful patches and generates language descriptions for target object classes within each selected patch. This transformation enables the datasets to be directly used for vision-language tasks, expanding their applicability.

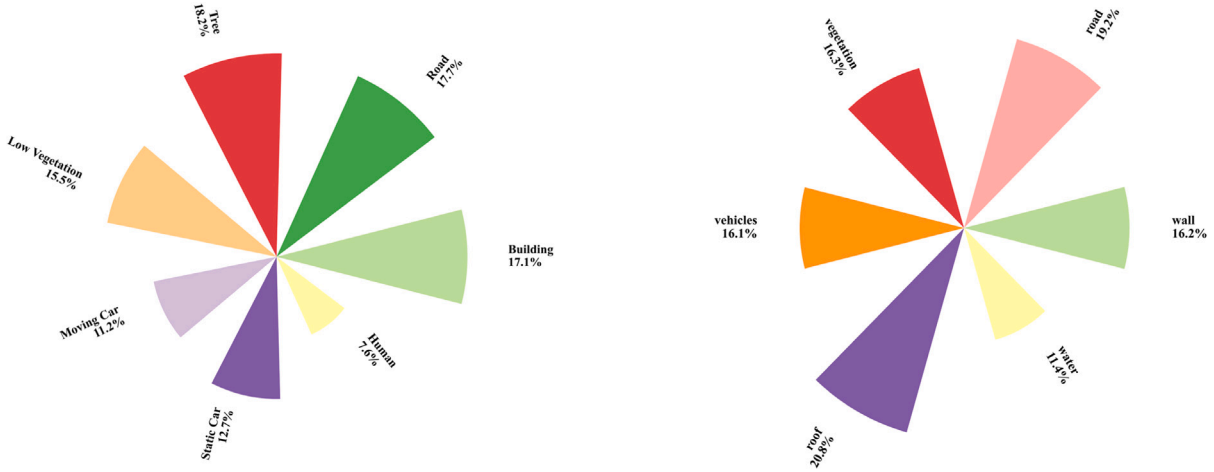
To gain insights into the linguistic and semantic characteristics of the generated dataset, we present a word cloud of the 100 most frequent words in Fig. 3, offering an overview of the variety of object descriptions. Additionally, Fig. 4 shows the image category distribution, providing a general understanding of the dataset composition and the occurrence of different object categories in UAVid-RIS and VDD-RIS.

3.2. Advantages and disadvantages of the designed pipeline

The designed pipeline offers several advantages that make it highly effective for UAV-RIS tasks, including:

- **Fully Automated Process.**
  The dataset generation pipeline is entirely automatic, eliminating the need for manual annotations. This significantly reduces human effort and makes it highly scalable for large-scale datasets.
- **Leverages Pre-existing Datasets.**
  Instead of requiring new annotations, the method takes advantage of already labelled segmentation datasets, making it efficient and cost-effective.
- **Multi-Label Representation.**
  A single image can have multiple referring expressions since it may contain multiple object categories, providing a richer semantic understanding and enabling a more comprehensive interpretation of the scene.
- **Scalable for Large Datasets.**
  The fully automatic pipeline allows for the generation of large-scale datasets without significant computational overhead, making it ideal for deep learning applications that require vast amounts of training data.
- **Diverse Language Descriptions.**
  Since the dataset's descriptions are generated by a large language model, it provides various expressions for the same object category. This enhances the dataset's linguistic diversity, making it more robust for vision-language models.

At the same time, there are also certain limitations that should be considered, including:



(a) Distribution of image categories of UAVid-RIS (b) Distribution of image categories of VDD-RIS

Fig. 4. Distribution of image categories of (a) UAVid-RIS and (b) VDD-RIS.

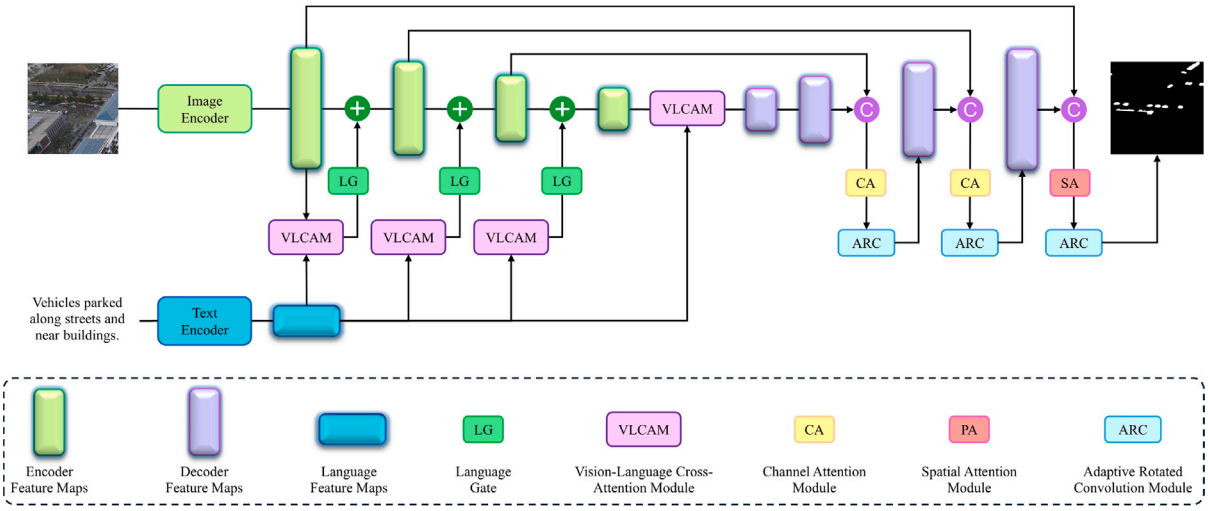


Fig. 5. Overview of the proposed AeroReformer.

- **Lack of Fine-Grained Object Features.**
  The descriptions generated by the language model focus on object categories rather than detailed attributes such as colour, texture, shape, or exact dimensions. This may limit the dataset's applicability for fine-grained vision tasks. Additionally, although semantically conflicting words are explicitly excluded in the prompt, the model may still occasionally include them in the generated text, leading to potential inconsistencies in the descriptions.
- **Limited Spatial and Positional Context.**
  While the dataset retains spatial and positional information, it is provided at the class level rather than for individual objects. Since the annotations correspond to entire object categories rather than distinct instances, precise localisation of single objects is not explicitly available.
- **Data Quality Depends on the Pre-Existing Dataset.**
  The overall quality of the generated dataset is inherently dependent on the accuracy and granularity of the segmentation dataset. If the segmentation labels are noisy, incomplete, or overly coarse, it may negatively impact the quality of the generated language descriptions and corresponding annotations.

Despite these limitations, the generated dataset provides a scalable and efficient solution for vision-language learning in UAV imagery, making it a valuable resource for automatic referring segmentation. As the first publicly available dataset for UAV-RIS, it establishes a foundational benchmark for future research in this field. All data generation code will be openly released to facilitate research and drive advancements in the remote sensing community.

4. Methodology

4.1. Problem formulation

This study aims to tackle the challenge of referring image segmentation in UAV imagery, where the objective is to generate an accurate segmentation mask for a target category based on a given natural language description. Formally, let $I \in \mathbb{R}^{H\times W\times C}$ denote an aerial image, where H, W, and C correspond to the image height, width, and number of channels, respectively. A textual query $T = \{t_1, t_2, \ldots, t_N\}$ serves as the semantic reference, where N represents the number of words or tokens in the description.

The goal is to predict a binary segmentation mask $O \in \{0,1\}^{H\times W}$, where each pixel $p \in I$ is classified as either belonging to the category described by T or not. Given a dataset $\Omega = \{(I_i, T_i, G_i)\}_{i=1}^{Num}$, where $G_i \in \{0,1\}^{H\times W}$ represents the corresponding ground truth mask and Num is the total number of samples, the objective is to develop a

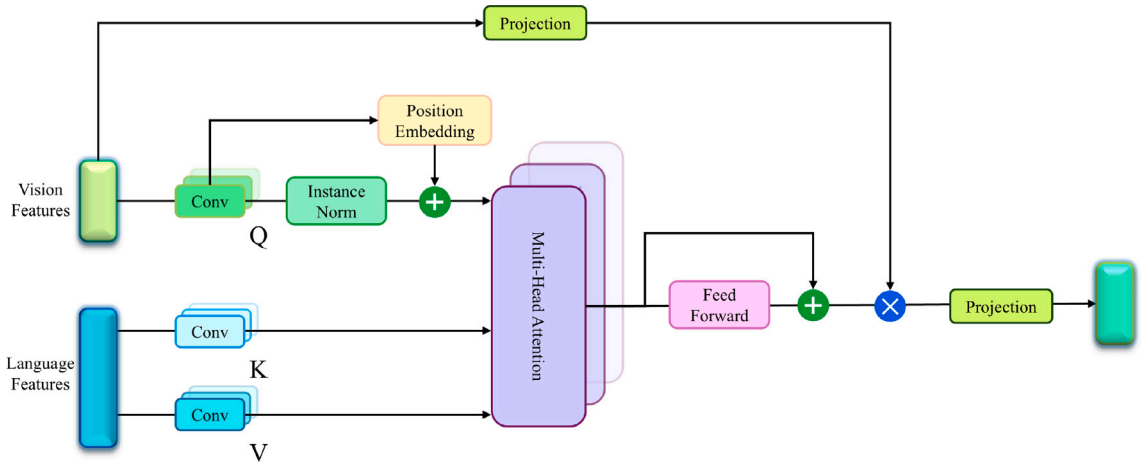


Fig. 6. Pipeline of the Vision-Language Cross-Attention Module (VLCAM).

function f that maps the image-text pair (I, T) to O by effectively learning cross-modal associations between linguistic descriptions and visual features.

4.2. Overview of the architecture

The overall architecture of our proposed AeroReformer is depicted in Fig. 5. Our AeroReformer builds upon LAVT (Yang et al., 2022), maintaining its encoders *e.g.,* Swin Transformer (Liu et al., 2021) and BERT (Devlin et al., 2019) for extracting multi-modal inputs while improving vision-language fusion and mask prediction. Meanwhile, we propose a vision-language cross-attention fusion module, enhancing the interaction between visual and linguistic features. Additionally, we introduce a rotation-aware multi-scale fusion decoder, allowing better adaptation to aerial imagery with varying orientations.

4.2.1. Vision-language cross-attention fusion module

To effectively integrate linguistic and visual information, we introduce a vision-language cross-attention fusion module that enhances cross-modal feature interaction, as shown in Fig. 6. This module replaces the Pixel-Word Attention Module (PWAM) with a more structured mechanism that utilises multi-head cross-attention to improve feature alignment between textual and visual representations.

Given an input aerial image $I \in \mathbb{R}^{H\times W\times C}$ and a natural language expression $T = \{t_1, t_2, \ldots, t_N\}$, our model extracts corresponding visual and linguistic features using hierarchical encoders. Let $F_V \in \mathbb{R}^{H'W'\times C_V}$ denote the visual features extracted from the image, where $H'W'$ represents the spatial dimension of the feature map and C_V is the number of visual feature channels. Similarly, let $F_L \in \mathbb{R}^{N\times C_L}$ represent the linguistic features extracted from the text, where N is the number of tokens, and C_L is the language feature dimension.

To facilitate cross-modal interactions between vision and language, we first project the features into a common embedding space using convolutional layers:

$$\begin{aligned} Q_V &= \mathscr{I}\left(\mathscr{C}_{1D}^{C_K}(F_V)\right), \\ K_L &= \mathscr{C}_{1D}^{C_K}(F_L), \\ V_L &= \mathscr{C}_{1D}^{C_V}(F_L). \end{aligned} \tag{1}$$

represents the 1D convolution operation, $\mathscr{I}$ denotes instance normalisation. C_K represents the key-query dimension, while C_V denotes the value dimension. The query features Q_V are extracted from vision features F_V using a 1D convolution followed by instance normalisation. The key K_L and value V_L are computed from the language features F_L using separate 1D convolutional layers. The positional encoding is enabled for query features:

$$Q_V = Q_V + P_V, \tag{2}$$

where P_V is a learnable positional encoding that provides spatial awareness to the vision features.

Next, we compute multi-head attention scores using scaled dot-product attention, while each head can be defined as:

$$\text{Attn} = \text{softmax}\left(\frac{Q_V K_L^T}{\sqrt{C_K}}\right), \tag{3}$$

where the dot product of queries and keys is scaled by $\sqrt{C_K}$ to stabilise gradients and prevent extreme values. The attended visual-language features are then computed as:

$$F_{VL} = \mathscr{I}\left(\mathscr{C}_{1D}(\text{Attn}\cdot V_L)\right). \tag{4}$$

Thereafter, a Feed-Forward Network (FFN) is applied, consisting of two convolutional layers with ReLU activation:

$$F_{FFN} = \mathscr{L}\left(\mathscr{C}_{1D}\left(\mathscr{D}(\mathscr{R}(\mathscr{C}_{1D}(F_{VL})))\right) + F_{VL}\right). \tag{5}$$

A residual connection is added between the input and output of the FFN to stabilise training and facilitate feature propagation. Specifically, the output of the FFN is element-wise added to the input before being passed through a subsequent layer normalisation operation $\mathscr{L}$. This mechanism follows the standard transformer design pattern and helps preserve important information while allowing the network to learn complex transformations. The cross-modal representation $F_{FFN} \in \mathbb{R}^{H'W'\times C_V}$ is then fused with the vision features via element-wise interaction:

$$F_{Fused} = F_{FFN} \odot \mathscr{P}(F_V), \tag{6}$$

where $\odot$ represents element-wise multiplication, generating enriched visual features that incorporate linguistic information. Projection $\mathscr{P}$ including convolution, GeLU activation and dropout is applied to vision features to stabilise distributions. Finally, the fused representation undergoes a final projection :

$$F_{out} = \mathscr{P}(F_{Fused}). \tag{7}$$

4.2.2. Rotation-aware multi-scale fusion decoder

To take full advantage of extracted features, the rotation-aware multi-scale fusion decoder is designed with multi-scale feature aggregation operations. Specifically, given a set of encoder feature maps at

different scales:

$$\begin{aligned} X_1 &\in \mathbb{R}^{B\times C_1\times \frac{H}{4}\times \frac{W}{4}}, \\ X_2 &\in \mathbb{R}^{B\times C_2\times \frac{H}{8}\times \frac{W}{8}}, \\ X_3 &\in \mathbb{R}^{B\times C_3\times \frac{H}{16}\times \frac{W}{16}}, \\ X_4 &\in \mathbb{R}^{B\times C_4\times \frac{H}{32}\times \frac{W}{32}}. \end{aligned} \tag{8}$$

where X_i represents the feature maps at different resolutions, and C_i denotes the corresponding number of channels. To maintain scale consistency, lateral transformations are applied using 1×1 convolutions first:

$$L_i = \mathscr{R}(\mathscr{C}_{2D}(X_i)), \quad i \in \{1,2,3,4\} \tag{9}$$

where $\mathscr{C}_{2D}$ is 2D convolutions.

Thereafter, upsampled decoder feature maps Y_i are concatenated with the corresponding lateral feature maps obtained from the encoder:

$$F_i = \text{Concat}(L_i, Y_i), \quad i \in \{1,2,3\} \tag{10}$$

This concatenated feature maps are then refined using either a channel attention module ($i = 2, 3$) or a spatial attention module ($i = 1$), which leverages L2 normalisation and adaptive weighting instead of traditional softmax-based dot-product attention, reducing both time and memory costs based on our previous work (Li et al., 2021a,b). The refined feature representation is given by:

$$F'_{\text{CA}} = X + \gamma_c \cdot \text{reshape}\left(\frac{\sum_n V_n + Q \odot (K \odot V)}{H\times W + Q \odot \sum_n K_n + \varepsilon}\right), \tag{11}$$

$$F'_{\text{SA}} = X + \gamma_s \cdot \text{reshape}\left(\frac{\sum_c V_c + Q \odot (K \odot V)}{H\times W + Q \odot \sum_c K_c + \varepsilon}\right), \tag{12}$$

where X is the input feature map, and Q, K, and V are the query, key, and value matrices, respectively, obtained from learned projections of X. The terms $\sum_n V_n$ and $\sum_c V_c$ represent aggregations of the value features across spatial and channel dimensions, respectively. ε is a small constant added for numerical stability. The scaling factors γ_c and γ_s are learnable parameters for the channel and spatial attention mechanisms. Here, $\odot$ denotes element-wise multiplication with broadcasting as necessary. The operations involving Q, K, and V follow batch-wise summations over intermediate dimensions, consistent with tensor contraction patterns implemented using PyTorch's `einsum` function. The "reshape" operation restores the feature tensor to its original spatial dimensions (H, W). Please refer to the code for the detailed implementation.

This formulation integrates both self-attention mechanisms, effectively capturing feature correlations across different dimensions while maintaining computational efficiency through L2-normalised weighting. Specifically, the channel attention module is applied to the first two fused feature maps, as they contain multiple channels representing hierarchical multi-scale information. This mechanism models dependencies between different channels, allowing the network to dynamically recalibrate inter-channel relationships and enhance contextual coherence. Meanwhile, for the final fused feature map, which is at the highest resolution and contains detailed spatial semantics, the spatial attention mechanism is employed. This ensures that long-range pixel dependencies are effectively captured, refining feature distributions across spatial dimensions.

To further improve feature representation across different orientations, which frequently occur in UAV imagery, we incorporate the Adaptive Rotated Convolution (ARC) module (Pu et al., 2023) into the fused features. Unlike standard convolution, where a fixed kernel is applied to all inputs, ARC adapts its filters to align with the directional variations present in imagery. Specifically, given an input feature map X, the routing function $\mathscr{F}$ predicts a set of rotation angles θ and corresponding weights λ:

$$\theta, \lambda = \mathscr{F}(X). \tag{13}$$

Each of the n convolution kernels W_i is then rotated according to its corresponding predicted angle:

$$W'_i = \text{Rotate}(W_i, \theta_i), \quad i = 1, 2, \ldots, n. \tag{14}$$

The rotated kernels are then used to convolve with the input feature map, and their outputs are combined in a weighted manner:

$$Y = \sum_{i=1}^{n} \lambda_i (W'_i * X). \tag{15}$$

This approach improves the model's ability to capture features from objects with varying orientations while maintaining computational efficiency.

5. Results and discussions

5.1. Experimental setting

5.1.1. Datasets

To evaluate the proposed method, we conducted extensive experiments on two newly developed UAV-RIS datasets, UAVid-RIS and VDD-RIS. Both datasets contain high-resolution images, all cropped to a size of 1024×1024 pixels. For both UAVid and VDD, we follow their official data splits when generating the RIS versions. All image patches are generated based directly on these splits to ensure consistency with the original datasets and fair benchmarking.

- UAVid-RIS. This dataset consists of 7035 images, divided into 3215 for training, 1163 for validation, and 2657 for testing. UAVid (Lyu et al., 2020) is designed for UAV-based scene understanding in complex urban environments, capturing both static and dynamic objects. The dataset features oblique-view aerial imagery with a camera angle of approximately 45 degrees, offering richer contextual information than nadir-view images. The data is collected from UAVs flying at an altitude of around 50 m, with high-resolution frames extracted from 4K video recordings. The dataset includes diverse street scenes with objects such as vehicles, pedestrians, buildings, roads, vegetation, billboards, and traffic infrastructure. To ensure meaningful patch selection during RIS generation, class-specific minimum area thresholds were applied: Building (15%), Tree (20%), Road (5%), Low Vegetation (10%), Moving Car (0.4%), Static Car (0.5%), and Human (0.1%).
- VDD-RIS. This dataset contains 1941 images, split into 1269 for training, 399 for validation, and 273 for testing. VDD (Cai et al., 2023) is collected across 23 locations in Nanjing, China, covering diverse environments, including urban, rural, industrial, and natural landscapes. The dataset incorporates variations in camera angles, with images captured at 30, 60, and 90 degrees (nadir view), allowing for more comprehensive scene representation. The drone altitude ranges from 50 to 120 m, ensuring a balance between scene complexity and fine-grained details. The dataset also introduces temporal and seasonal diversity, with images taken at different times of the day and in different seasons. For patch filtering, class-specific minimum area thresholds were set as follows: Wall (8%), Road (10%), Vegetation (50%), Vehicle (0.5%), Roof (20%), and Water (5%).

5.1.2. Implementation details

We implemented our method in PyTorch (Paszke et al., 2019), utilising the pre-trained base BERT (Devlin et al., 2019) for language processing and the Swin Transformer (Liu et al., 2021) initialised with ImageNet-22K (Deng et al., 2009) weights for visual encoding.

All images were resized to 480×480 pixels, and no data augmentation (e.g., rotation, flipping) was applied. Training was conducted with a batch size of 8 for 40 epochs on UAVid-RIS and 10 epochs on VDD-RIS using the AdamW optimiser (Loshchilov and Hutter, 2017) with a weight decay of 0.01 and an initial learning rate of 0.0005. Following the baseline LAVT (Yang et al., 2022), cross-entropy loss was used for optimisation. All experiments were performed on an NVIDIA RTX 5000 Ada GPU.

Table 1
Performance comparison of different methods on UAVid-RIS. The table includes Precision at different IoU thresholds (Pr@0.5 to Pr@0.9), mean Intersection over Union (mIoU), overall Intersection over Union (oIoU), and the visual and textual encoders used in each method.

Method	Visual encoder	Textual encoder	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
RIS-DMMI	ResNet-101	BERT	79.75	70.19	54.08	36.13	9.26	67.10	76.76
LAVT	Swin-B	BERT	84.45	75.75	59.60	38.71	10.91	69.32	78.76
LGCE	Swin-B	BERT	83.97	74.60	59.69	39.37	11.33	69.52	79.06
RMSIN	Swin-B	BERT	85.71	77.91	64.06	46.52	17.76	72.05	81.10
ASDA	CLIP-ViT-B	CLIP	78.55	69.51	56.38	37.67	10.05	67.17	76.59
MAFN	Swin-B	BERT	85.10	77.76	63.68	44.26	16.33	71.64	80.67
AeroReformer	Swin-B	BERT	**86.34**	**79.07**	**65.60**	**47.12**	**18.52**	**72.79**	**81.53**

Table 2
Performance comparison of different methods on VDD-RIS. The table includes Precision at different IoU thresholds (Pr@0.5 to Pr@0.9), mean Intersection over Union (mIoU), and overall Intersection over Union (oIoU), and the visual and textual encoders used in each method.

Method	Visual encoder	Textual encoder	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
RIS-DMMI	ResNet-101	BERT	87.08	77.86	62.73	46.13	23.25	72.66	78.06
LAVT	Swin-B	BERT	90.04	86.72	76.01	56.09	34.32	77.80	82.51
LGCE	Swin-B	BERT	89.67	84.13	73.80	53.51	32.47	76.78	82.09
RMSIN	Swin-B	BERT	91.14	85.98	76.38	57.20	35.06	78.22	83.58
ASDA	CLIP-ViT-B	CLIP	90.77	85.98	76.01	57.56	35.79	78.13	81.99
MAFN	Swin-B	BERT	91.51	87.82	79.70	60.89	35.06	79.21	83.97
AeroReformer	Swin-B	BERT	**92.99**	**89.30**	**81.55**	**63.47**	**38.75**	**80.72**	**85.38**

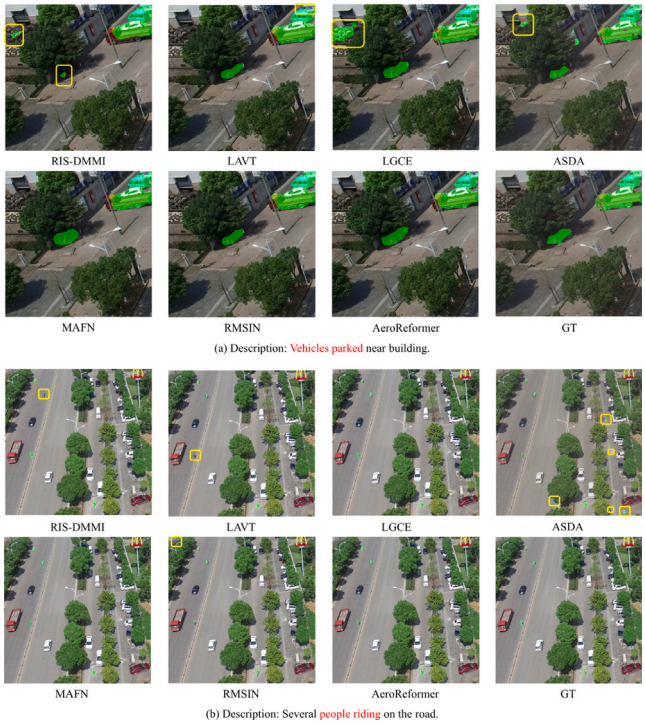


Fig. 7. Visual comparison of referring segmentation results on UAVid-RIS dataset.

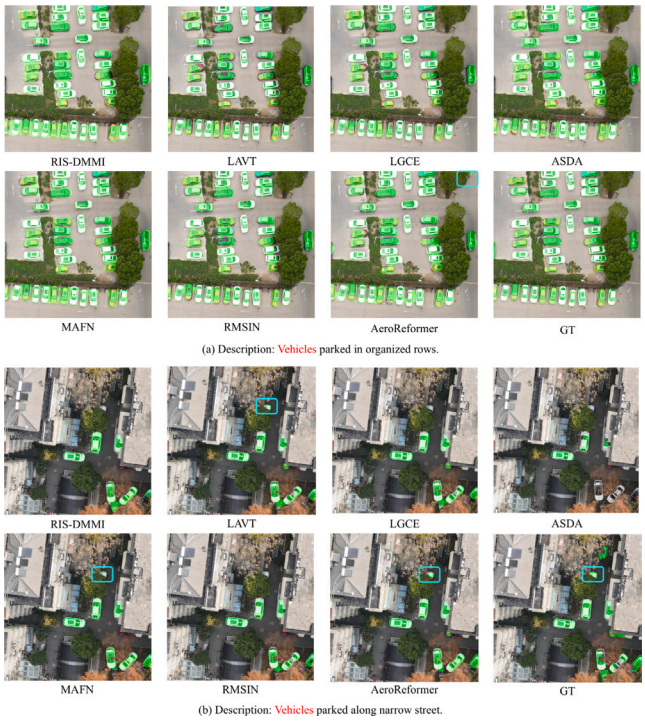


Fig. 8. Visual comparison of referring segmentation results on VDD-RIS dataset.

5.1.3. Model evaluation

For a fair comparison with previous methods (Yuan et al., 2024; Liu et al., 2024; Yang et al., 2022), we adopted the same evaluation metrics, including mean Intersection over Union (mIoU), overall Intersection over Union (oIoU), and Precision at the 0.5, 0.6, 0.7, 0.8 and 0.9 IoU thresholds (Pr@X).

The mIoU measures the average IoU between predicted and ground-truth masks across all test samples, giving equal weight to both large and small objects. In contrast, oIoU favours large objects by computing the ratio of the total intersection area to the total union area across all test samples. Additionally, Pr@X evaluates model performance at different IoU thresholds, reflecting the proportion of successfully predicted samples at each level. Higher values for these metrics indicate better segmentation performance.

In addition to these metrics, we also report the class-wise IoU and class-wise mIoU to provide a more detailed analysis of segmentation performance across different object categories. Unlike object-based mIoU, which is calculated per test sample, class-wise mIoU is computed per class rather than per object, ensuring that the evaluation captures category-level segmentation accuracy instead of object-instance accuracy.

Table 3
Class-wise Intersection over Union (IoU) and mean IoU (mIoU) for different methods on UAVid-RIS.

Method	Building	Road	Tree	Low vegetation	Moving car	Static car	Human	mIoU
RIS-DMMI	84.48	77.91	79.67	64.08	67.04	50.18	24.11	63.92
LAVT	86.87	78.68	80.91	67.95	68.98	53.53	23.75	65.81
LGCE	87.58	79.36	80.67	67.97	67.78	58.07	24.01	66.49
RMSIN	89.02	84.33	82.01	69.23	**72.31**	54.79	24.72	68.06
ASDA	84.16	80.29	79.81	63.77	59.47	39.22	20.86	61.08
MAFN	88.66	83.52	81.67	68.54	71.28	59.29	**26.78**	68.53
AeroReformer	**89.12**	**84.82**	**82.35**	**69.45**	72.14	**60.58**	26.68	**69.31**

Table 4
Class-wise Intersection over Union (IoU) and mean IoU (mIoU) for different methods on VDD-RIS.

Method	Wall	Road	Vegetation	Vehicles	Roof	Water	mIoU
RIS-DMMI	65.95	77.21	89.43	66.38	76.77	80.63	76.06
LAVT	72.62	82.94	90.04	69.07	81.01	91.30	81.16
LGCE	70.63	80.75	89.97	69.42	82.09	90.03	80.48
RMSIN	73.16	82.64	89.39	70.11	84.88	90.54	81.79
ASDA	72.05	**83.72**	84.16	60.15	84.84	**92.81**	79.62
MAFN	74.68	82.53	89.70	70.25	84.88	90.94	82.16
AeroReformer	**76.82**	82.57	**91.78**	**70.74**	**86.21**	91.25	**83.23**

5.2. UAV referring image segmentation performance

In this section, we evaluate the performance of seven different referring image segmentation methods, including RIS-DMMI (Hu et al., 2023), LAVT (Yang et al., 2022), LGCE (Yuan et al., 2024), RMSIN (Liu et al., 2024), ASDA (Yue et al., 2024), MAFN (Shi and Zhang, 2025) and the proposed AeroReformer.

5.2.1. Quantitative results

Overall Performance: As shown in Tables 1 and 2, the proposed AeroReformer consistently achieves the highest scores across all evaluation metrics, demonstrating its superior segmentation capability. In UAVid-RIS, AeroReformer achieves an mIoU of 72.79 and an oIoU of 81.53, surpassing the second-best method, RMSIN, by 0.74 in mIoU and 0.43 in oIoU. Similarly, on VDD-RIS, AeroReformer achieves an mIoU of 80.72 and an oIoU of 85.38, outperforming MAFN by 1.51 in mIoU and 1.41 in oIoU. These improvements highlight AeroReformer's effectiveness in capturing fine-grained segmentation details across different datasets.

Precision at Different IoU Thresholds: In terms of precision at varying IoU thresholds, Tables 1 and 2 illustrate that AeroReformer consistently outperforms the second-best method across all threshold levels. On UAVid-RIS, AeroReformer achieves the highest Pr@0.5 score of 86.34, surpassing RMSIN by 0.63, and maintains its lead at Pr@0.9 with 18.52, exceeding RMSIN by 0.76. On VDD-RIS, AeroReformer achieves a Pr@0.5 of 92.99, improving upon MAFN by 1.48, and maintains the best performance at Pr@0.9 with 38.75, surpassing ASDA by 2.96. These improvements confirm AeroReformer's robustness and reliability in maintaining segmentation accuracy under different IoU thresholds.

Class-wise IoU Analysis: A deeper analysis of class-wise IoU scores in Tables 3 and 4 further supports AeroReformer's superior performance. On UAVid-RIS, AeroReformer achieves the highest IoU scores in five out of seven categories. Compared to the second-best method, MAFN, AeroReformer improves static car segmentation by 1.29. MAFN slightly outperforms AeroReformer in the human category, but AeroReformer still maintains the highest overall mIoU. On VDD-RIS, AeroReformer achieves the highest IoU scores in four out of six categories. It surpasses second best by 2.14 in wall segmentation and 0.49 in vehicle segmentation. LAVT outperforms AeroReformer in the water category, while ASDA performs best in the road category. Despite this, AeroReformer achieves the highest overall mIoU of 83.23, improving by 1.07.

5.2.2. Qualitative results

To further evaluate the segmentation performance of different RIS methods, we present qualitative comparisons on UAVid-RIS and VDD-RIS in Figs. 7 and 8. Each example consists of segmentation results from seven different methods: RIS-DMMI, LAVT, LGCE, RMSIN, ASDA, MAFN and the proposed AeroReformer, along with the ground truth (GT). The visualised results highlight AeroReformer's ability to produce more precise and contextually accurate segmentations.

Results on UAVid-RIS: Fig. 7 illustrates segmentation results for two different referring expressions: *"Vehicles parked near building"* and *"Several people riding on the road"*. In the first example, RIS-DMMI, LGCE, and ASDA incorrectly classify certain materials in the top-left storage yard as vehicles, resulting in inaccurate segmentation. In contrast, AeroReformer produces consistent segmentation results that closely align with the ground truth. In the second example, which involves detecting people riding on the road, RIS-DMMI and LAVT fail to identify all relevant targets, while RMSIN and ASDA mistakenly classify road patches as people. AeroReformer accurately segments the riding individuals without misclassifying unrelated elements. These results highlight AeroReformer's effectiveness in distinguishing small objects within UAV imagery.

Results on VDD-RIS: Fig. 8 presents segmentation results for two different referring expressions: *"Vehicles parked in organised rows"* and *"Vehicles parked along a narrow street"*. In the first example, all methods perform well in detecting parked vehicles. Notably, in the top right corner, a black car is partially covered under the trees; even though the ground truth does not label it, AeroReformer successfully detects it, demonstrating its superior ability to capture occluded objects. In the second example, which depicts vehicles parked along a narrow street, RIS-DMMI, LGCE, ASDA, and RMSIN fail to distinguish a parked car from surrounding trees, particularly when a car is partially covered by foliage. AeroReformer accurately differentiate the vehicles, minimising misclassification. This demonstrates AeroReformer's strong performance in complex urban environments with occlusions and varying object scales.

Summary: The qualitative comparisons across UAVid-RIS and VDD-RIS confirm that AeroReformer produces more precise and contextually accurate segmentations than existing methods. It consistently outperforms the second-best method by correctly distinguishing between similar objects and capturing finer details. These results validate AeroReformer's effectiveness in complex aerial scenes with dynamic and static objects.

5.2.3. Generalisation experiments

To further evaluate the generalisability of the proposed AeroReformer, we conduct two additional experiments focusing on language robustness and domain transfer:

Cross-LLM Evaluation: To assess the model's robustness to language variation, we evaluate all models on descriptions generated by a different MLLM, Llama-3.2-11B-Vision (Grattafiori et al., 2024) using the identical prompt. The model is still trained on UAVid-RIS using descriptions generated by Qwen2.5-VL-7B (Bai et al., 2025). This setup simulates practical deployment scenarios where referring expressions may vary depending on the language model used. As shown in Table 5, AeroReformer maintains strong performance under this shift in language domain, achieving the highest scores across all metrics, including

Table 5
Performance comparison of different methods on UAVid-RIS using description generated by Llama. The table includes Precision at different IoU thresholds (Pr@0.5 to Pr@0.9), mean Intersection over Union (mIoU), overall Intersection over Union (oIoU), and the visual and textual encoders used in each method.

Method	Visual encoder	Textual encoder	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
RIS-DMMI	ResNet-101	BERT	73.81	63.98	50.17	33.04	8.05	62.24	74.09
LAVT	Swin-B	BERT	73.32	62.29	45.62	27.29	5.68	60.05	69.90
LGCE	Swin-B	BERT	70.76	60.71	45.43	26.95	6.10	58.98	71.87
RMSIN	Swin-B	BERT	75.20	68.87	57.43	41.29	15.13	63.19	78.40
ASDA	CLIP-ViT-B	CLIP	51.41	46.26	38.01	26.38	7.68	44.16	42.91
MAFN	Swin-B	BERT	77.72	69.89	56.57	39.48	13.32	65.56	77.33
AeroReformer	Swin-B	BERT	**82.80**	**75.76**	**62.93**	**45.35**	**16.79**	**69.77**	**80.07**

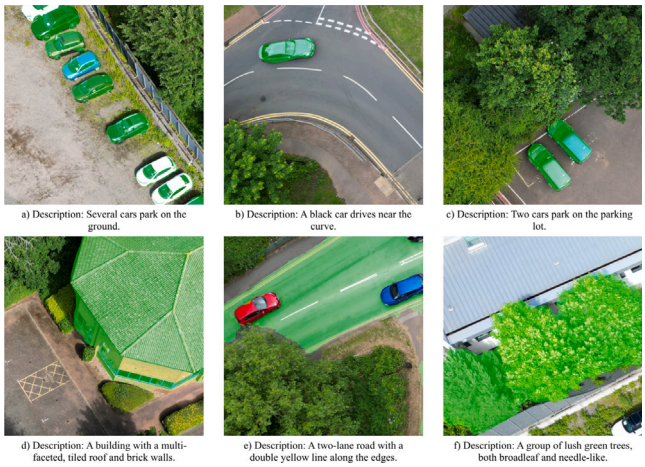


Fig. 9. Model generation ability test on UAV images captured at the University of Warwick.

a mIoU of 69.77 and Pr@0.5 of 82.80. Despite the noticeable performance drop across all models compared to their original test setting, AeroReformer demonstrates the best generalisation, outperforming the next-best method (MAFN) by +4.21 in mIoU and +5.08 in Pr@0.5. This suggests that the proposed AeroReformer is more resilient to shifts in linguistic style and semantics, even without explicit re-training. ASDA, which uses CLIP-based encoders, performs poorly under this variation, likely due to limited language-text alignment adaptation in aerial contexts. Overall, these results affirm AeroReformer's robustness in handling diverse linguistic inputs, a desirable trait for real-world UAV-RIS applications.

Cross-Location Testing: We test the generalisation ability of AeroReformer on UAV images captured on the University of Warwick campus. These images differ significantly from UAVid-RIS in terms of geographic location, flight altitude, and camera angle. This experiment demonstrates the model's adaptability to diverse environmental and acquisition conditions without retraining or fine-tuning. Specifically, the model accurately segments parked vehicles under occlusions (Fig. 9a,c), moving cars on curved roads (Fig. 9b), and detailed structures such as multi-faceted buildings (Fig. 9d). It also successfully highlights fine-grained objects like two-lane roads (Fig. 9e) and distinguishes vegetation types even under diverse lighting and texture conditions (Fig. 9f). These qualitative results confirm AeroReformer's strong generalisation capacity to novel geographic settings and unseen flight parameters.

The results from both experiments highlight the flexibility of the proposed approach and its potential for real-world deployment in varied settings.

5.3. Ablation study

To analyse the contribution of each proposed module in AeroReformer, we conduct an ablation study on the VDD-RIS dataset. Table 6 presents the results when replacing the Rotation-Aware Multi-Scale Fusion (RAMSF) decoder and the Vision-Language Cross Attention Module (VLCAM). The evaluation is performed using Pr@0.5, Pr@0.7, Pr@0.9 and mIoU.

Table 6
Ablation study results on the VDD-RIS dataset. The table evaluates the impact of RAMSF and VLCAM, showing Precision at IoU thresholds Pr@0.5, Pr@0.7, and Pr@0.9, and mean Intersection over Union (mIoU). A checkmark (✓) indicates the inclusion of a module.

RAMSF	VLCAM	Pr@0.5	Pr@0.7	Pr@0.9	mIoU
–	–	89.30	73.06	28.04	76.09
–	✓	90.04	78.60	36.53	78.27
✓	–	91.14	77.86	36.16	78.83
✓	✓	92.99	81.55	38.75	80.72

Baseline: For the baseline configuration (i.e., without RAMSF and VLCAM), we retain the overall model structure while replacing the VLCAM module with a sentence feature vector globally pooled from all words. As for the RAMSF, in the absence of a decoder, we replace it with the module from LAVT (Yang et al., 2022). As expected, the performance drops notably across all metrics, achieving 89.30 in Pr@0.5, 73.06 in Pr@0.7, 28.04 in Pr@0.9, and 76.09 in mIoU. These results demonstrate the limited capacity of the backbone alone in handling fine-grained and cross-modal reasoning tasks.

VLCAM: With VLCAM included, improvements of +0.74 in Pr@0.5, +5.54 in Pr@0.7, +8.49 in Pr@0.9, and +2.18 in mIoU are achieved compared to the baseline. This indicates that VLCAM significantly enhances vision-language interaction, particularly under stricter IoU thresholds.

RAMSF: When RAMSF is included, gains of +1.84 in Pr@0.5, +4.80 in Pr@0.7, +8.12 in Pr@0.9, and +2.74 in mIoU are observed. These improvements confirm the effectiveness of RAMSF in multi-scale spatial feature fusion and detail preservation.

AeroReformer: The full model, with both RAMSF and VLCAM included, achieves the best overall performance: 92.99 in Pr@0.5, 81.55 in Pr@0.7, 38.75 in Pr@0.9, and 80.72 in mIoU. This represents improvements of +3.69 (Pr@0.5), +8.49 (Pr@0.7), +10.71 (Pr@0.9), and +4.63 (mIoU) over the baseline. These results clearly demonstrate the complementary benefits of RAMSF and VLCAM.

6. Conclusions

In this work, we proposed a fully automated dataset construction pipeline that transforms pre-existing UAV segmentation datasets into referring segmentation benchmarks. The designed pipeline leverages segmentation masks and large language models to generate diverse and contextually accurate referring expressions. This method was applied to UAVid and VDD, producing UAVid-RIS and VDD-RIS, two novel datasets that expand the applicability of vision-language segmentation in UAV imagery. In addition, we introduced AeroReformer, a novel framework for referring image segmentation that integrates the Rotation-Aware Multi-Scale Fusion (RAMSF) decoder and the Vision-Language Cross-Attention Module (VLCAM) to enhance spatial feature fusion and cross-modal alignment. AeroReformer consistently

outperforms comparative methods on UAVid-RIS and VDD-RIS, demonstrating superior segmentation accuracy in challenging UAV environments with occlusions, scale variations, and fine-grained object details. The ablation study further validates the necessity of RAMSF and VL-CAM, showing that their combination significantly boosts segmentation performance.

While AeroReformer achieves significant improvements, it still relies on a separate vision encoder for segmentation. Future work should explore the integration of segmentation capabilities directly into large language models (LLMs), eliminating the need for external vision modules.

CRediT authorship contribution statement

Rui Li: Formal analysis, Data curation, Conceptualisation, Investigation, Methodology, Project administration, Software, Validation, Visualisation, Writing - original draft. **Xiaowei Zhao:** Conceptualisation, Funding acquisition, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has received funding from the UK Engineering and Physical Sciences Research Council (grant number: EP/Y016297/1).

Data availability

The UAVid dataset is available at https://uavid.nl/ and the VDD dataset is available at https://huggingface.co/datasets/RussRobin/VDD. The code to generate the UAVid-RIS and VDD-RIS is available at https://github.com/lironui/AeroReformer, and the generated text references are available at https://huggingface.co/datasets/lironui/UAVid-RIS and https://huggingface.co/datasets/lironui/VDD-RIS.

References

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al., 2025. Qwen2.5-VL. Technical Report, arXiv preprint arXiv:2502.13923.

Cai, W., Jin, K., Hou, J., Guo, C., Wu, L., Yang, W., 2023. Vdd: Varied drone dataset for semantic segmentation. arXiv preprint arXiv:2305.13608.

Chen, K., Zhang, J., Liu, C., Zou, Z., Shi, Z., 2025. RSRefSeg: Referring remote sensing image segmentation with foundation models. arXiv preprint arXiv:2501.06809.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186.

Ding, H., Liu, C., Wang, S., Jiang, X., 2022. VLT: Vision-language transformer and query generation for referring segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 45 (6), 7900–7916.

Ding, Y., Xu, H., Wang, D., Li, K., Tian, Y., 2024. Visual selection and multi-stage reasoning for rsvg. IEEE Geosci. Remote. Sens. Lett..

Dong, Z., Sun, Y., Gu, Y., Liu, T., 2024. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. arXiv preprint arXiv:2410.08613.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H., 2020. Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4424–4433.

Hu, R., Rohrbach, M., Darrell, T., 2016. Segmentation from natural language expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 108–124.

Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., Luo, P., 2023. Beyond one-to-one: Rethinking the referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4067–4077.

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J., 2024. Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589.

Lei, S., Xiao, X., Zhang, T., Li, H.-C., Shi, Z., Zhu, Q., 2024. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens..

Li, R., Li, K., Kuo, Y.-C., Shu, M., Qi, X., Shen, X., Jia, J., 2018. Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5745–5753.

Li, K., Vosselman, G., Yang, M.Y., 2025. Scale-wise bidirectional alignment network for referring remote sensing image segmentation. arXiv preprint arXiv:2501.00851.

Li, K., Wang, D., Xu, H., Zhong, H., Wang, C., 2024. Language-guided progressive attention for visual grounding in remote sensing images. IEEE Trans. Geosci. Remote Sens..

Li, R., Zhao, X., 2024. LSwinSR: UAV imagery super-resolution based on linear swin transformer. IEEE Trans. Geosci. Remote Sens..

Li, R., Zheng, S., Duan, C., Su, J., Zhang, C., 2021a. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P.M., 2021b. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 181, 84–98.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, C., Ding, H., Jiang, X., 2023. Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23592–23601.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A., 2017. Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1271–1280.

Liu, S., Ma, Y., Zhang, X., Wang, H., Ji, J., Sun, X., Ji, R., 2024. Rotated multi-scale interaction network for referring remote sensing image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26658–26668.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M.Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. ISPRS J. Photogramm. Remote Sens. 165, 108–119.

Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P., 2018. Dynamic multimodal instance segmentation guided by natural language queries. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 630–645.

Mienna, I.M., Klanderud, K., Ørka, H.O., Bryn, A., Bollandsås, O.M., 2022. Land cover classification of treeline ecotones along a 1100 km latitudinal transect using spectral-and three-dimensional information from UAV-based aerial imagery. Remote. Sens. Ecol. Conserv. 8 (4), 536–550.

Nagaraja, V.K., Morariu, V.I., Davis, L.S., 2016. Modeling context between objects for referring expression understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, pp. 792–807.

Pan, Y., Sun, R., Wang, Y., Zhang, T., Zhang, Y., 2024. Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2031–2040.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Pu, Y., Wang, Y., Xia, Z., Han, Y., Wang, Y., Gan, W., Wang, Z., Song, S., Huang, G., 2023. Adaptive rotated convolution for rotated object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6589–6600.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Shao, H., Song, P., Mu, B., Tian, G., Chen, Q., He, R., Kim, G., 2021. Assessing city-scale green roof development potential using Unmanned Aerial Vehicle (UAV) imagery. Urban For. Urban Green. 57, 126954.

Sharma, R., Arya, R., 2022. UAV based long range environment monitoring system with industry 5.0 perspectives for smart city infrastructure. Comput. Ind. Eng. 168, 108066.

Shi, H., Li, H., Meng, F., Wu, Q., 2018. Key-word-aware network for referring expression image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 38–54.

Shi, L., Zhang, J., 2025. Multimodal-aware fusion network for referring remote sensing image segmentation. IEEE Geosci. Remote. Sens. Lett..

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sun, Y., Feng, S., Li, X., Ye, Y., Kang, J., Huang, X., 2022. Visual grounding in remote sensing images. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 404–412.

Tahir, M.N., Lan, Y., Zhang, Y., Wenjiang, H., Wang, Y., Naqvi, S.M.Z.A., 2023. Application of unmanned aerial vehicles in precision agriculture. In: Precision Agriculture. Elsevier, pp. 55–70.

Trepekli, K., Balstrøm, T., Friborg, T., Fog, B., Allotey, A.N., Kofie, R.Y., Møller-Jensen, L., 2022. UAV-borne, LiDAR-based elevation modelling: A method for improving local-scale urban flood risk assessment. Nat. Hazards 113 (1), 423–451.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.

Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H., 2022. Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165.

Ye, L., Rochan, M., Liu, Z., Wang, Y., 2019. Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10502–10511.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L., 2018. Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1307–1315.

Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L., 2016. Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, pp. 69–85.

Yuan, Z., Mou, L., Hua, Y., Zhu, X.X., 2024. Rrsis: Referring remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens..

Yue, P., Lin, J., Zhang, S., Hu, J., Lu, Y., Niu, H., Ding, H., Zhang, Y., Jiang, G., Cao, L., et al., 2024. Adaptive selection based referring image segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1101–1110.

Zhan, Y., Xiong, Z., Yuan, Y., 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE Trans. Geosci. Remote Sens. 61, 1–13.

Zhang, T., Wen, Z., Kong, B., Liu, K., Zhang, Y., Zhuang, P., Li, J., 2025. Referring remote sensing image segmentation via bidirectional alignment guided joint prediction. arXiv preprint arXiv:2502.08486.

Zhang, Q., Zheng, S., Zhang, C., Wang, X., Li, R., 2023. Efficient large-scale oblique image matching based on cascade hashing and match data scheduling. Pattern Recognit. 138, 109442.

Zhao, R., Shi, Z., Zou, Z., 2021. High-resolution remote sensing image captioning based on structured attention. IEEE Trans. Geosci. Remote Sens. 60, 1–14.